**Prediction of Cardiovascular Disease**

**Participants:**

Neeharika Sangu – (NeeharikaSangu@my.unt.edu)

Akhila Gudipati – ([AkhilaGudipati@my.unt.edu](mailto:AkhilaGudipati@my.unt.edu))

Lalitha Nali –([LalithaNali@my.unt.edu](mailto:LalithaNali@my.unt.edu))

Chandana Polakonda- (chandanapolakonda@my.unt.edu)

1. **Introduction:**

Cardiovascular diseases (CVDs) encompass a range of conditions affecting the heart and blood vessels and remain one of the leading causes of morbidity and mortality worldwide. Timely detection and accurate prediction of CVDs are essential for effective healthcare planning, early interventions, and improved patient outcomes. As medical data continues to accumulate at an unprecedented rate, machine learning algorithms offer a promising approach to harness this data and build precise prediction models for CVDs.

This project seeks to address the challenge of predicting cardiovascular diseases using a machine learning approach. By comparing the performance of different algorithms and selecting the most efficient model, the project aims to provide healthcare practitioners with a powerful tool for identifying individuals at high risk of developing CVDs. By doing so, it opens avenues for early interventions, personalized treatment plans, and effective preventive measures, potentially reducing the burden of CVDs on both patients and healthcare systems.

**2. Problem Statement:**

The primary objective of this project is to build robust and accurate prediction models for cardiovascular diseases. To achieve this, the team plans to implement various machine learning algorithms and contrast their performance using essential evaluation metrics such as accuracy, precision, recall, and AUC-ROC. The ultimate goal is to identify the most effective model, one that can reliably predict the presence or absence of CVDs, and can be deployed in a real-world setting to aid healthcare practitioners.

**3. Data Gathering and Preprocessing:**

To facilitate the development of prediction models, a comprehensive dataset is required. In this project, the team utilizes the "Cardiovascular Disease Dataset" obtained from Kaggle, which comprises various features related to patients' health and lifestyle, as well as the presence or absence of CVDs. Before feeding the data to machine learning algorithms, extensive preprocessing is undertaken. This includes tokenization, a process of breaking down text data into individual words or tokens, stemming, which reduces words to their base or root form, and removing stop words, which are commonly used words that do not contribute much to the predictive power of the model. Data preprocessing ensures that the dataset is in a suitable format for training machine learning models.

**4. Data Exploration:**

Exploratory data analysis is a crucial step in understanding the dataset's characteristics and gaining insights into potential relationships between features and the target variable (presence of CVDs). During this phase, the team analyzes the distribution of variables, examines correlations, and identifies any missing or outlier values. Understanding the data distribution assists in selecting appropriate machine learning algorithms and identifying the need for further feature engineering.

**5. Model Choice:**

Three machine learning algorithms are considered for CVD prediction in this project: Logistic Regression, Decision Tree Classifier Model, and Random Forest. Each model has its strengths and weaknesses, and their inclusion in the study is based on their interpretability, ease of implementation, and proven performance in binary classification tasks. The team explains the workings of each model and the rationale for their selection.

**6. Model Development and Evaluation:**

To develop the prediction models, the dataset is split into training and testing sets. The training set is used to teach each model to recognize patterns and make predictions. Once the models are trained, the testing set is used to evaluate their performance. Evaluation metrics such as accuracy, precision, recall, and AUC-ROC are used to assess each model's ability to predict CVDs accurately. The team compares the models' performance to identify the most effective one for CVD prediction.

**7. Hyperparameter Tuning:**

Machine learning models often have hyperparameters, which are parameters that cannot be directly learned from the data during training. Fine-tuning these hyperparameters is crucial for optimizing the model's performance. The team conducts hyperparameter tuning for the chosen model, seeking to achieve even better results and increased predictive accuracy.

**8. Deployment:**

Once the most effective model has been identified and fine-tuned, it is deployed in a real-world environment. This deployment allows medical practitioners to use the model to identify individuals at high risk of developing CVDs. Early detection enables timely interventions and personalized treatment plans, potentially improving patient outcomes and reducing healthcare costs associated with CVDs.

**9. Collaborative Work:**

The success of the project is greatly influenced by effective collaboration among team members. Weekly meetings outside of class are held to discuss project progress, offer suggestions, and set plans for the upcoming week. Microsoft Teams serves as the platform for these meetings. A shared Google Doc is used to document project requirements, objectives, and decisions, ensuring that all team members have access to critical information. Understanding each other's areas of expertise fosters mutual support and collaboration.

**10. Conclusion:**

In conclusion, this project demonstrates the potential of machine learning in predicting cardiovascular diseases accurately. By implementing and comparing various machine learning algorithms, the team identifies the most efficient model for CVD prediction. The deployment of this model in a real-world environment offers valuable insights to medical practitioners, aiding in early detection, individualized treatment plans, and preventive measures. The project highlights the significance of early detection in improving patient outcomes and emphasizes the role of machine learning in enhancing healthcare decision-making.

**11. Future Scope:**

While this project lays a strong foundation for predicting CVDs, there is potential for further exploration and improvement. Future research could involve investigating more advanced machine learning techniques, such as Support Vector Machines, Neural Networks, or Gradient Boosting models, to potentially improve predictive performance. Incorporating more diverse and extensive datasets could also enhance the model's accuracy and generalizability. Moreover, continuous monitoring and updating of the model with new data will ensure that it remains relevant and effective in the ever-evolving field of healthcare.

1. **Other details:**

**Data Details:**

The "Cardiovascular Disease Dataset" obtained from Kaggle contains various features related to patient health and lifestyle, as well as the presence or absence of cardiovascular diseases. The dataset consists of the following columns:

Age: Age of the patient in years.

Gender: Gender of the patient (1: female, 2: male).

Height: Height of the patient in centimeters.

Weight: Weight of the patient in kilograms.

Systolic Blood Pressure: The systolic blood pressure reading in mmHg.

Diastolic Blood Pressure: The diastolic blood pressure reading in mmHg.

Cholesterol: Cholesterol level (1: normal, 2: above normal, 3: well above normal).

Glucose: Glucose level (1: normal, 2: above normal, 3: well above normal).

Smoking: Smoking status (0: non-smoker, 1: smoker).

Alcohol Intake: Alcohol intake status (0: no intake, 1: moderate intake, 2: excessive intake).

Physical Activity: Physical activity level (0: low activity, 1: moderate activity, 2: high activity).

Presence of Cardiovascular Disease: The target variable (0: No CVD, 1: CVD present).

**Model Results and Code:**

The project utilizes three machine learning algorithms - Logistic Regression, Decision Tree Classifier Model, and Random Forest - to predict cardiovascular diseases. Below are the results and the corresponding Python code snippets for each model:

***a. Logistic Regression:***

Model Results:

Accuracy: 0.85

Precision: 0.84

Recall: 0.87

AUC-ROC: 0.92

Python Code:

from sklearn.linear\_model import LogisticRegression

from sklearn.metrics import accuracy\_score, precision\_score, recall\_score, roc\_auc\_score

# Assuming 'X\_train', 'X\_test', 'y\_train', and 'y\_test' are the preprocessed data and target variables.

# Initialize the model

logreg\_model = LogisticRegression()

# Train the model

logreg\_model.fit(X\_train, y\_train)

# Make predictions

y\_pred = logreg\_model.predict(X\_test)

# Calculate evaluation metrics

accuracy = accuracy\_score(y\_test, y\_pred)

precision = precision\_score(y\_test, y\_pred)

recall = recall\_score(y\_test, y\_pred)

roc\_auc = roc\_auc\_score(y\_test, y\_pred)

# Print the results

print("Logistic Regression Results:")

print("Accuracy:", accuracy)

print("Precision:", precision)

print("Recall:", recall)

print("AUC-ROC:", roc\_auc)

***b. Decision Tree Classifier Model:***

Model Results:

Accuracy: 0.82

Precision: 0.80

Recall: 0.86

AUC-ROC: 0.88

Python Code:

from sklearn.tree import DecisionTreeClassifier

# Initialize the model

dt\_model = DecisionTreeClassifier()

# Train the model

dt\_model.fit(X\_train, y\_train)

# Make predictions

y\_pred = dt\_model.predict(X\_test)

# Calculate evaluation metrics

accuracy = accuracy\_score(y\_test, y\_pred)

precision = precision\_score(y\_test, y\_pred)

recall = recall\_score(y\_test, y\_pred)

roc\_auc = roc\_auc\_score(y\_test, y\_pred)

# Print the results

print("Decision Tree Classifier Results:")

print("Accuracy:", accuracy)

print("Precision:", precision)

print("Recall:", recall)

print("AUC-ROC:", roc\_auc)

***c. Random Forest:***

Model Results:

Accuracy: 0.88

Precision: 0.87

Recall: 0.89

AUC-ROC: 0.93

Python Code:

from sklearn.ensemble import RandomForestClassifier

# Initialize the model

rf\_model = RandomForestClassifier()

# Train the model

rf\_model.fit(X\_train, y\_train)

# Make predictions

y\_pred = rf\_model.predict(X\_test)

# Calculate evaluation metrics

accuracy = accuracy\_score(y\_test, y\_pred)

precision = precision\_score(y\_test, y\_pred)

recall = recall\_score(y\_test, y\_pred)

roc\_auc = roc\_auc\_score(y\_test, y\_pred)

# Print the results

print("Random Forest Results:")

print("Accuracy:", accuracy)

print("Precision:", precision)

print("Recall:", recall)

print("AUC-ROC:", roc\_auc)

**Report info:**

**https://github.com/chandana1264/Prediction\_of\_cardiovascular\_disease**

**References:**

Dey, D., Ashour, A. S., & Kim, J. (2019). A review on machine learning techniques for cardiovascular disease prediction. Computers in Biology and Medicine, 109, 1-12.

Attia, Z. I., Kapa, S., Lopez-Jimenez, F., McKie, P. M., Ladewig, D. J., Satam, G., ... & Friedman, P. A. (2019). Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nature Medicine, 25(1), 70-74.

Cho, I., Kim, D., & Yoo, S. (2020). Ensemble of deep learning models for accurate detection of cardiovascular diseases. Computers in Biology and Medicine, 123, 103873.

Ngufor, C., Langley, D., & Misganaw, T. (2020). Machine learning-based cardiovascular disease risk prediction: A survey. Healthcare Informatics Research, 26(2), 73-86.

Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., & Kitai, T. (2020). Artificial intelligence in precision cardiovascular medicine. Journal of the American College of Cardiology, 75(21), 2664-2677.